Kullback-Leibler Cluster Entropy: an information measure of genome complexity

Anna Carbone
Politecnico di Torino
Torino, Italy
anna.carbone@polito.it

Renato Ferrero
Dipartimento di Automatica e Informatica
Politecnico di Torino
Turin, Italy 1234-5678-9012

Filippo Gandino
Dipartimento di Automatica e Informatica
Politecnico di Torino
Turin, Italy emailfilippo.gandino@polito.it

Chiara Panico
Dipartimento di Automatica e Informatica
Politecnico di Torino
Turin, Italy 1234-5678-9012

Abstract

Long memory processes yield important information about how complex systems evolve under positive or negative correlation among the individual processes and components. Rescaled range analysis (R/S), detrended fluctuation analysis (DFA), detrended moving average analysis (DMA) and generalized Hurst exponent (GHE) are among widely adopted methods to discriminate between correlated and anticorrelated sequences in terms of Hurst exponent. The recently proposed Kullback-Leibler cluster entropy [1, 2], a measure of divergence between probability distributions of long-range correlated sequences, has the computational advantage of not relying on a linear regression of log-log data plots to estimate the Hurst exponent. The Shannon cluster entropy has been deployed on the 24 chromosomes of the Hg-19 human reference genome [3] and to quantify heterogeneity of human pangenome minigraphs [4]. The Kullback-Leibler cluster entropy has been adopted for comparing the T2T-CHM13+Y human reference genome including centromeric regions and short arms, i.e. gaps still present in the previously published reference assembly [5] and for genes involved in epigenetic regulation, neurodevelopmental disorders and other complex diseases [6]. The cluster entropy can accurately detect the recently added structurally complex nucleotide regions in the 24 chromosomes, through the local variability of their correlation exponents. To test the cluster entropy performance, the results are referred to the nucleotide positions of the earlier assembly (GRCh38) and prove the algorithm accuracy in detecting the newly added strands. Each nucleotide sequence is first mapped into a numerical representation, then its statistical divergence is quantified against synthetic fractional Brownian motions with correlation degrees assigned as input parameter used as reference models. By comparing the entropy-based profiles across sequences, we identify characteristic patterns that reflect intrinsic structural properties. Importantly, the detected features remain stable under different segmentation schemes implemented through either overlapping or non-overlapping sliding window computational strategies. Overall, our findings demonstrate the robustness of the proposed information-theoretical measure and support to extend its application to comparative genomics, structural genome analysis, and functional gene characterization. Biological implications are discussed, and potential uses of the Kullback-Leibler cluster divergence in machine-learning and deep-learning pipelines for genomic data are outlined.

CCS Concepts

• **Theory of computation → Randomness, geometry and discrete structures**; • **Mathematics of computing → Information theory**; • **Applied computing → Life and medical sciences**.

ACM Reference Format:
Anna Carbone, Renato Ferrero, Filippo Gandino, and Chiara Panico. 2026. Kullback-Leibler Cluster Entropy: an information measure of genome complexity. In . ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

References

[1] Anna Carbone and Linda Ponta. Relative cluster entropy for power-law correlated sequences. *SciPost Physics*, 13(3):076, 2022.
[2] Linda Ponta and Anna Carbone. Kullback-Leibler cluster entropy to quantify volatility correlation and risk diversity. *Physical Review E*, 142(3):205–210, 2025.
[3] Anna Carbone. Information measure for long-range correlated sequences: the case of the 24 human chromosomes. *Scientific reports*, 3(1):2721, 2013.
[4] Renato Ferrero, Filippo Gandino, and Anna Carbone. Information theoretic clustering of the human pangenome minigraph. *Pattern Recognition Letters*, 191:117–123, 2025.
[5] Filippo Gandino, Renato Ferrero, and Anna Carbone. Kullback-Leibler cluster entropy: a complexity measure of the 24 chromosomes of the T2T-CHM13+Y human reference genome. *Available at SSRN 6308958*.
[6] Filippo Gandino, Chiara Panico, Renato Ferrero, Mieye Ombe, and Anna Carbone. Kullback-Leibler cluster entropy to quantify local correlation in human genes. Proc. of the Intn. Conf. on E-Health and Bioengineering-EHB 2025, 2026. in press.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn